Learning Unsupervised Word Mapping via Maximum Mean Discrepancy

Pengcheng Yang[1,2], Fuli Luo[2], Shuangzhi Wu[3], Jingjing Xu[2], and Dongdong Zhang[3]

[1] Center for Data Science, Beijing Institute of Big Data Research, Peking University
[2] MOE Key Lab of Computational Linguistics, School of EECS, Peking University
[3] Microsoft Research Asia
{yang_pc,luofuli,jingjingxu}@pku.edu.cn
{v-shuawu,dozhang}@microsoft.com

Abstract. Cross-lingual word embeddings aim at capturing common linguistic regularities of different languages. Recently, it has been shown that these embeddings can be effectively learned by aligning two disjoint monolingual vector spaces through a simple linear transformation (word mapping). In this work, we focus on learning such a word mapping without any supervision signal. Most previous work of this task adopts adversarial training or parametric metrics to perform distribution-matching, which typically requires a sophisticated alternate optimization process, either in the form of *minmax game* or intermediate *density estimation*. This alternate optimization process is relatively hard and unstable. In order to avoid such sophisticated alternate optimization, we propose to learn unsupervised word mapping by directly minimize the maximum mean discrepancy between the distribution of the transferred embedding and target embedding. Extensive experimental results show that our proposed model can substantially outperform several state-of-the-art unsupervised systems, and even achieves competitive performance compared to supervised methods. Further analysis demonstrates the effectiveness of our approach in improving stability.

Keywords: Cross-lingual · Embeddings · Unsupervised learning.

1 Introduction

It has been shown that word embeddings are capable of capturing meaningful representations of words [7]. Recently, more and more efforts turn to cross-lingual word embeddings, which benefit various downstream tasks ranging from unsupervised machine translation to transfer learning.

Based on the observation that the monolingual word embeddings share similar geometric properties across languages [19], an underlying idea is to align two disjoint monolingual vector spaces through a linear transformation. [23] further empirically demonstrates that the results can be improved by constraining the desired linear transformation as an orthogonal matrix, which is also proved theoretically by [22].

Recently, increasing effort has been motivated to learn word mapping without any supervision signal. One line of research focuses on designing heuristics [16] or utilizing structural similarity of monolingual embeddings [1,6,14]. However, these methods often require a large number of random restarts or additional skills such as reweighting [5] to achieve satisfactory results. Another line of research strives to learn

unsupervised word mapping by directly matching the distribution of the transferred embedding and target embedding. For instance, [8,17,25] implement the word mapping as the generator in the generative adversarial network (GAN), which is essentially a *minmax game*. [26] and [24] adopt the Earth Mover's distance and Sinkhorn distance as the optimized distance metrics respectively, both of which require intermediate *density estimation*. Although this line exhibits relatively excellent performance, both the *minmax game* and intermediate *density estimation* require alternate optimization. However, such a sophisticated alternate optimization process tends to cause a hard and unstable optimization problem [11].

In order to avoid the sophisticated alternate optimization process required by *minmax game* or intermediate *density estimation*, in this paper, we propose to learn unsupervised word mapping between different languages by directly minimize the *maximum mean discrepancy* (MMD) [12] between the distribution of the transferred embedding and target embedding. The MMD distance is a non-parametric metric, which measures the difference between two distributions. Compared to other parametric metrics, it does not require any intermediate *density estimation* as well as adversarial training. This MMD-based distribution-matching at one-step results in a relatively simple and stable optimization problem, which leads to improvements in the model performance.

The main contributions of this paper are summarized as follows:

- We propose to learn unsupervised word mapping by directly minimize maximum mean discrepancy between distribution of transferred embedding and target embedding, which avoids a relatively sophisticated alternate optimization process.
- Extensive experimental results show that our approach achieves better performance than several state-of-the-art unsupervised systems, and even achieves competitive performance compared to supervised methods. Further analysis demonstrates the effectiveness of our approach in improving stability.

2 Background

Here we briefly introduce the background knowledge of learning cross-lingual word embeddings based on the linear mapping between two monolingual embedding spaces. Let $\mathcal{X} = \{x_i\}_{i=1}^{n}$ and $\mathcal{Y} = \{y_i\}_{i=1}^{m}$ be two sets of n and m pre-trained monolingual word embeddings, which come from the source and target language, respectively. Our goal is to learn a word mapping $\mathbf{W} \in \mathbb{R}^{d \times d}$ so that for any source word embedding $x \in \mathbb{R}^d$, $\mathbf{W}x$ lies close to the embedding $y \in \mathbb{R}^d$ of its corresponding target language translation. Here d represents the dimension of pre-trained monolingual word embeddings. Furthermore, [22,23] show that the model performance can be improved by constraining the linear transformation $\mathbf{W}$ as an orthogonal matrix.

2.1 Supervised Scenarios

Suppose $\mathbf{X} \in \mathbb{R}^{n \times d}$ and $\mathbf{Y} \in \mathbb{R}^{n \times d}$ be the aligned monolingual word embedding matrices between two different languages, which means that $(\mathbf{X}_i, \mathbf{Y}_i)$ is the embedding

of the aligned word pair. Here $\mathbf{X}_i$ and $\mathbf{Y}_i$ denote the i-th row of $\mathbf{X}$ and $\mathbf{Y}$, respectively. Then, the optimal linear mapping $\mathbf{W}^*$ can be recovered by solving the following optimization problem:

$$\mathbf{W}^* = \underset{\mathbf{W} \in \mathcal{O}_d}{\operatorname{argmin}} ||\mathbf{X}\mathbf{W} - \mathbf{Y}||_{\mathrm{F}} \tag{1}$$

where $\mathcal{O}_d$ is the space composed of all $d \times d$ orthogonal matrices and $|| \cdot ||_{\mathrm{F}}$ refers to the Frobenius norm. Under the constraint of orthogonality of $\mathbf{W}$, Eq. (2) boils down to the Procrustes problem, which advantageously offers a closed form solution:

$$\mathbf{W}^* = \mathbf{U}\mathbf{V}^\top \tag{2}$$

where $\mathbf{U}\mathbf{S}\mathbf{V}^\top$ is the singular value decomposition of $\mathbf{X}^\top\mathbf{Y}$.

2.2 Unsupervised Scenarios

When involving in unsupervised cross-lingual embedding, one representative line of research focuses on learning the linear mapping $\mathbf{W}$ by matching the distribution of transferred embedding and target embedding. In other words, the optimal liner mapping $\mathbf{W}^*$ can be learned by making the distribution of $\mathbf{W}\mathcal{X}$ and $\mathcal{Y}$ as close as possible:

$$\mathbf{W}^* = \underset{\mathbf{W} \in \mathcal{O}_d}{\operatorname{argmin}} \textbf{Dist}(\mathcal{P}, \mathcal{Q}) \tag{3}$$

where $\mathcal{P}$ and $\mathcal{Q}$ denote distribution of the transferred embedding and target embedding, respectively. **Dist**$(\cdot, \cdot)$ is the optimized distance metric between two distributions, which can be adopted as *Jensen-Shannon Divergence* [17,8,25], *Wasserstein Distance* [26], *Sinkhorn Distance* [24], *Gromov Distance* [2], and so on.

3 Proposed Method

The most crucial component of our approach is MMD-matching. In addition, the iterative training and model initialization also play an important role in improving results. We elaborate on these three components in detail as follows.

3.1 MMD-Matching

In order to avoid sophisticated alternate optimization process required by adversarial training or intermediate *density estimation*, we directly minimize the maximum mean discrepancy between the distribution of the transferred embedding and target embedding. *Maximum mean discrepancy* (MMD) is a non-parametric metric that measures the difference between two distributions. It does not require any intermediate *density estimation* as well as adversarial training, thus avoiding a relative sophisticated alternate optimization.

Same as Section 2.2, we use $\mathcal{P}$ and $\mathcal{Q}$ to represent the distribution of the transferred embedding $\mathbf{W}\mathcal{X}$ and target embedding $\mathcal{Y}$, respectively, i.e., $\mathbf{W}x \sim \mathcal{P}$ and $y \sim \mathcal{Q}$.

Then, the difference between the distributions $\mathcal{P}$ and $\mathcal{Q}$ can be characterized by the MMD distance between $\mathcal{P}$ and $\mathcal{Q}$:

$$\mathrm{MMD}(\mathcal{P}, \mathcal{Q}) = \sup_{f \in \mathcal{F}} \left[\mathbb{E}_{\mathbf{W}x \sim \mathcal{P}} f(\mathbf{W}x) - \mathbb{E}_{y \sim \mathcal{Q}} f(y) \right] \tag{4}$$

where $\mathcal{F}$ is generally defined as a unit ball in *Reproducing Kernel Hilbert Space* (RKHS) $\mathcal{H}$. MMD applies a class of functions as a collection of trials to measure the difference between two distributions. Intuitively, for two similar distributions, the expectation of multiple trials should be close. $\mathrm{MMD}(\mathcal{P}, \mathcal{Q})$ in Eq. (4) reaches its minimum only when the distribution $\mathcal{P}$ and $\mathcal{Q}$ match exactly. Therefore, in order to match the distribution of transferred embedding and target embedding as exactly as possible, the optimal mapping $\mathbf{W}^*$ can be learned by solving the following optimization problem:

$$\min_{\mathbf{W} \in \mathcal{O}_d} \mathrm{MMD}(\mathcal{P}, \mathcal{Q}) \tag{5}$$

By means of *kernel trick* [13], the MMD distance between the distributions $\mathcal{P}$ and $\mathcal{Q}$ can be calculated as follows:

$$\begin{aligned} \mathrm{MMD}^2(\mathcal{P}, \mathcal{Q}) = & \mathbb{E}_{\mathbf{W}x \sim \mathcal{P}, \mathbf{W}x' \sim \mathcal{P}}[k(\mathbf{W}x, \mathbf{W}x')] \\ & + \mathbb{E}_{y \sim \mathcal{Q}, y' \sim \mathcal{Q}}[k(y, y')] \\ & - 2\mathbb{E}_{\mathbf{W}x \sim \mathcal{P}, y \sim \mathcal{Q}}[k(\mathbf{W}x, y)] \end{aligned} \tag{6}$$

where $k(\cdot, \cdot) : \mathbb{R}^d \times \mathbb{R}^d \mapsto \mathbb{R}$ is the kernel function in the RKHS space, such as polynomial kernel or Gaussian kernel. Due to the large size of search space (monolingual embedding space), it is intractable to directly calculate Eq. (6). Therefore, at the training stage, Eq. (6) can be estimated by the sampling method, which is formulated as:

$$\begin{aligned} \mathrm{MMD}^2(\mathcal{P}, \mathcal{Q}) = & \frac{1}{b^2} \Bigg\{ \sum_{i=1}^{b} \sum_{j=1}^{b} k(\mathbf{W}x_i, \mathbf{W}x_j) \\ & - 2 \sum_{i=1}^{b} \sum_{j=1}^{b} k(\mathbf{W}x_i, y_j) + \sum_{i=1}^{b} \sum_{j=1}^{b} k(y_i, y_j) \Bigg\} \end{aligned} \tag{7}$$

where b refers to the size of mini-batch.

Previous work [22,23] has shown that imposing the orthogonal constraint to the linear mapping $\mathbf{W}$ can lead to better performance. The orthogonal transformation not only preserves the quality of the monolingual embeddings, but also guarantees the consistency of Euclidean distance and the dot product of vectors. Therefore, in order to maintain the orthogonality of $\mathbf{W}$ during the training phase, we adopt the same update strategy proposed in [17]. In detail, after updating the linear mapping $\mathbf{W}$ with a certain optimizer in each learning step, we replace the original update of the matrix $\mathbf{W}$ with the following update rule:

$$\mathbf{W} \leftarrow (1 + \beta)\mathbf{W} - \beta(\mathbf{W}\mathbf{W}^T)\mathbf{W} \tag{8}$$

where β is a hyper-parameter. The results show that the matrix $\mathbf{W}$ is capable of staying close to the manifold of orthogonal matrices[4] after each update.

[4] In the experiment, we can observe that the eigenvalues of the matrix $\mathbf{W}$ all have a modulus close to 1.

Algorithm 1 The training process of our approach.

Require: source monolingual embeddings $\mathcal{X} = \{x_i\}_{i=1}^{n}$ and target monolingual embeddings $\mathcal{Y} = \{y_i\}_{i=1}^{m}$

Initialization:
  Utilize the structural similarity of embeddings to learn the initial word mapping $\mathbf{W}_0$
MMD-Matching:
  Randomly sample a batch of x from $\mathcal{X}$
  Randomly sample a batch of y from $\mathcal{Y}$
  Compress x and y to a lower feature space via Eq. (9)
  Compute the estimated MMD distance via Eq. (7)
  Update all model parameters via backward propagation
  Orthogonalize linear mapping $\mathbf{W}$ via Eq. (8)
Iterative Refinement:
  Repeat the following process:
    Build the pseudo-parallel dictionary $\mathbf{D}$ via Eq. (10)
    Learn a better $\mathbf{W}$ by solving Procrustes problem
  Until convergence

3.2 Compressing Network

At the training stage, Eq. (6) is estimated by the sampling method. The bias of estimation directly determines the accuracy of the calculation of the MMD distance. A reliable estimation of Eq. (6) generally requires the size of the mini-batch to be proportional to the dimension of the word embedding. Therefore, we adopt a *compressing network*[5] to map all embeddings into a lower feature space. Experimental results show that the use of *compressing network* can not only improve the performance of the model, but also provide significant computational savings. In detail, we implement the *compressing network* as a multilayer perceptron, which is formulated as follows:

$$\mathrm{CPS}(e) = \mathbf{W}_2\big(\max(0, \mathbf{W}_1 e + b_1)\big) + b_2 \tag{9}$$

where e refers to the input embedding and $\mathrm{CPS}(\cdot)$ represents the compressing network. $\mathbf{W}_1, \mathbf{W}_2, b_1$ and b_2 are learnable parameters.

3.3 Iterative Refinement and Initialization

Previous work has shown that refinement can bring a significant improvement in the quality of learned word mapping [6,17]. Therefore, after the optimization process of matching the distribution $\mathcal{P}$ and $\mathcal{Q}$ based on the MMD distance converges, we apply the iterative refinement to further improve results. For each source word s, we apply the currently learned linear mapping $\mathbf{W}$ to find its nearest target translation $\hat{t}$ based on the cosine similarity to build the pseudo-parallel dictionary $\mathbf{D} = \{(s, \hat{t})\}$. Formally,

$$\hat{t} = \underset{t}{\operatorname{argmax}} \cos(\mathbf{W}x_s, y_t) \tag{10}$$

[5] We train a specific compression network separately for each language pair.

where x_s and y_t represent the pre-trained embedding of the source word s and target word t, respectively. Subsequently, we apply the Procrustes solution in Eq. (2) on the pseudo-parallel dictionary to learn a better word mapping. As a result, the improved word mapping is able to induce a more accurate bilingual dictionary, which in turn helps to learn better word mapping. The two tasks of inducing bilingual dictionary and learning word mapping can be boosted with each other iteratively.

Another important issue is the initialization of model parameters. Considering that an inappropriate initialization tends to cause the model to stuck in poor local optimum [1,24,26], following previous work [1,26], we provide a warm-start for the proposed MMD-matching. Specifically, we take advantage of the structural similarity of embeddings to construct a pseudo-parallel dictionary, and then obtain the initial word mapping $\mathbf{W}_0$ by solving the Procrustes problem. Readers can refer to [6] for the detailed approach.

In summary, at the training stage, we first utilize the structural similarity of embeddings to obtain the initialized word mapping $\mathbf{W}_0$. Then, we perform MMD-matching to match the distribution of transferred embedding and target embedding. Finally, iterative refinement is adopted to further improve model performance. An overview of the training process is summarized in Algorithm 1.

4 Experiments

4.1 Evaluation Tasks

Following previous work [17,24], we evaluate our proposed model on bilingual lexicon induction. The goal of this task is to retrieve the translation of given source word. We use the bilingual lexicon constructed by [17]. Here we report accuracy with *nearest neighbor retrieval* based on cosine similarity[6].

4.2 Baselines

We compare our approach with the following supervised and unsupervised methods.

Supervised baselines. [19] proposes to learn the desired linear mapping by minimizing mean squared error. [23] normalizes the word vectors on a hypersphere and constrains the linear transform as an orthogonal matrix. [21] tries to alleviate the hubness problem by optimizing the inverse mapping. [27] refines the model in an unsupervised manner by initializing and regularizing it to be close to the direct transfer model. [3] proposes a generalized framework including orthogonal mapping and length normalization. [4] presents a self-learning framework to improve model performance.

Unsupervised baselines. [25] implements the word mapping as the generator in the GAN and [26] goes a step further to apply Wasserstein GAN by minimizing the earth mover's distance. [17] presents the cross-domain similarity local scaling (CSLS). [24] incorporates the Sinkhorn distance as a distributional similarity measure, and jointly learns the word embedding transfer in both directions.

[6] We also tried CSLS retrieval and results show that our approach achieved consistent improvement over baselines. Due to page limitations, we only report results with cosine similarity.

Methods	DE-EN	EN-DE	ES-EN	EN-ES	FR-EN	EN-FR	IT-EN	EN-IT
Supervised:								
[19]	61.93	**73.07**	74.00	**80.73**	71.33	**82.20**	68.93	**77.60**
[23]	67.73	69.53	77.20	78.60	76.33	78.67	72.00	73.33
[21]	**71.07**	63.73	**81.07**	74.53	**79.93**	73.13	**76.47**	68.13
[27]	67.67	69.87	77.27	78.53	76.07	78.20	72.40	73.40
[3]	69.13	72.13	78.27	80.07	77.73	79.20	73.60	74.47
[4]	68.07	69.20	75.60	78.20	74.47	77.67	70.53	71.67
Unsupervised:								
[25]	40.13	41.27	58.80	60.93	-	57.60	43.60	44.53
[26]	-	55.20	70.87	71.40	-	-	64.87	65.27
[17]	69.73	71.33	79.07	78.80	77.87	78.13	74.47	75.33
[24]	67.00	69.33	77.80	79.53	75.47	77.93	72.60	73.47
Ours	**70.33***	**71.53***	**79.33***	**79.93***	**78.87***	**78.40***	**74.73***	**75.53***

Table 1: Results of different methods on bilingual lexicon induction. **Bold** indicates the best supervised and unsupervised results, respectively. "-" means that the model fails to converge and hence the result is omitted. "*" indicates that our model is significantly better than the best performing unsupervised baseline. Language codes: EN=English, DE=German, ES=Spanish, FR=French, IT=Italian.

4.3 Experiment Settings

We use publicly available 300-dimensional *fastText* word embeddings. The size of the parameter matrices $\mathbf{W}_1$ and $\mathbf{W}_2$ in the *compressing network* are [300, 1024] and [1024, 50], respectively. The batch size is set to 1280 and β in Eq.(8) is set to 0.01. We use a mixture of 10 isotropic Gaussian (RBF) kernels with different bandwidths σ as in [18]. We use the Adam optimizer with initial learning rate 10^{-5}. We adopt the unsupervised criterion proposed in [17] as both an early-stopping criterion and a model selection criterion. For a fair comparison, we apply the same initialization and iterative refinement to all baselines.

5 Results and Discussion

In this section, we report all experimental results and conduct in-depth analysis.

5.1 Experimental Results

The experimental results of our approach and all baselines are shown in Table 1. Results show that our proposed model can achieve better performance than all unsupervised baselines on all test language pairs. Compared to the supervised methods, it is gratifying that our approach also achieves completely comparable performance. This demonstrates that the use of MMD is of great help to improve the quality of the word mapping. Our approach adopts a non-parametric metric that does not require intermediate *density estimation* or adversarial training. This enables the matching process of

Models	EN-ES	EN-FR	EN-DE	EN-IT
[25]	0.28	0.36	0.51	0.37
[26]	0.41	0.42	0.71	0.36
[17]	0.26	0.28	0.43	0.29
[24]	0.49	0.61	0.67	0.54
Ours	**0.21**	**0.27**	**0.35**	**0.24**

Table 2: Standard deviation (%) of the accuracy of 10 repeated experiments. The language codes are shown in Table 1.

the distribution of transferred embedding and target embedding to avoid sophisticated alternate optimization, leading to the improvements in the model performance.

5.2 Effectiveness of Improving Stability

Most of the previous work requires sophisticated alternate optimization, resulting in a relatively hard and unstable training process. This poor stability also leads to a large variance in the model performance. In order to verify that our proposed model based on the MMD metric can do a great favor to improving the stability, we repeat 10 sets of experiments on the bilingual lexicon induction task with different random seeds and calculate the standard deviation of the accuracy of these 10 sets of experiments. Table 2 presents the relevant results[7].

As shown in Table 2, the baselines suffer from poor stability in the repeated experiments. The variance of the accuracy of the baseline [26] reaches 0.71% in the EN-DE language pair. In contrast, our approach is able to achieve an obvious decline in standard deviation, which means a significant improvement in stability. For instance, the standard deviation on the EN-DE language pair is dropped from 0.43% to 0.35%, which powerfully illustrates the effectiveness of our approach in improving stability. With the MMD metric, our approach is able to perform distribution-matching in one step. This avoids the trade-off between the two optimization problems in the alternate optimization, resulting in a significant improvement in stability.

5.3 Effectiveness of Improving Distant Language Pairs

Previous work has shown that learning word mapping between distant language pairs remains an intractable challenge. Distant languages exhibit huge differences in both grammar and syntax, leading that their embedding spaces have different structures. Surprisingly, our approach can substantially outperform baselines on distant language pairs, as shown in Table 3. For instance, on the EN-ZH language pair, our method beats the best result of baselines by a margin of 2.4%.

Existing methods require sophisticated alternate optimization, whose performance depends on a delicate balance between two optimization procedures during training.

[7] Due to page limitations, for each language pair, we only show results in one direction because the conclusions drawn from the other direction are the same. For example, we only show EN-FR and ignore FR-EN. Same in Table 3, Table 4, and Figure 1.

Models	EN-BG	EN-CA	EN-SV	EN-ZH
[25]	-	17.87	-	18.07
[26]	16.47	29.33	-	22.73
[17]	22.53	35.60	32.80	26.07
[24]	25.07	40.53	38.47	29.87
Ours	**27.13**	**42.47**	**39.93**	**32.27**

Table 3: Performance of different methods on four distant language pairs. Language codes: EN=English, BG=Bulgarian, CA=Catalan, SV=Swedish, ZH=Chinese.

Models	EN-ES	EN-FR	EN-DE	EN-IT
Full model	79.93	78.40	71.53	75.53
w/o Compression	76.87	75.93	70.73	73.47
w/o MMD-matching	71.60	72.53	68.20	71.40
w/o Refinement	55.80	65.27	61.00	58.67
w/o Initialization	-	-	-	-

Table 4: Ablation study on the bilingual lexicon induction task. "-" means that the model fails to converge and hence the result is omitted. The language codes are shown in Table 1.

Once this training balance is not well maintained, the model performance tends to degrade. For instance, GAN [25] is vulnerable to *mode collapse* when learning word mapping between distant languages. For embedding spaces of a distant language pair, some subspaces are similar between two languages, while others show language-specific structures that are hard to align. Since it is easy for the generator to obtain high rewards on the former subspaces from discriminator, the generator is encouraged to optimize on the former subspaces and ignores the latter ones, which results in a poor alignment model on language-specific dissimilar subspaces. In contrast, our approach bypasses this issue by avoiding alternate optimization, which reduces the strict requirements for the training balance. The MMD distance strives to directly align the global embedding spaces of the two languages via kernel functions, which models the dissimilar embedding subspace of distant language pairs more effectively, leading to better performance.

5.4 Ablation Study

In order to understand the importance of different components of our approach, here we perform an ablation study by training multiple versions of our model with some missing components. The relevant results are presented in Table 4.

According to Table 4, the most critical component is initialization, without which the proposed model will fail to converge. The reason is that an inappropriate initialization tends to cause the model to stuck in a poor local optimum. The same initialization sensitivity issue is also observed by [1,26,24]. This sensitivity issue is ingrained and difficult to eliminate. In addition, as shown in Table 4, the final refinement can bring a significant improvement in the model performance. What we need to emphasize is that although the missing of MMD-matching brings the relatively weak decline in model

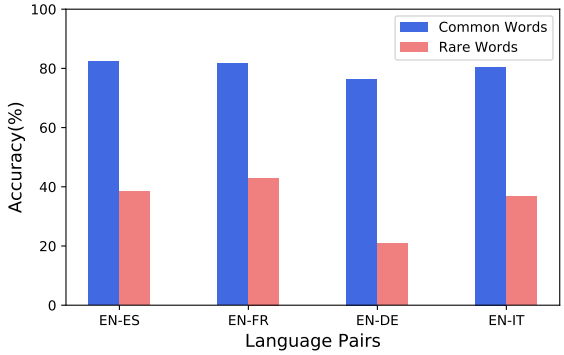

Fig. 1: The performance of our approach in common words and rare words on the bilingual lexicon induction task. Common words are the most frequent 20,000 words, and the remaining are regarded as rare words.

performance, it is still a key component to guide the model to learn a better final word mapping. For instance, with the help of MMD-matching, the accuracy increases from 71.60% to 79.93% on the EN-ES testing pair. Our approach is able to avoid sophisticated alternating optimization, leading to an improvement in the model performance. In addition, the results also show that the compressing network also plays an active role in improving accuracy. The compressing network aims to project the embedding into a lower feature space, making the estimation of the MMD distance more accurate.

5.5 Error Analysis

In the experiment, we find that all methods exhibit relatively poor performance when translating rare words on the bilingual lexicon induction task. Figure 1 shows the performance of our approach on the common word pairs and the rare word pairs, from which we can see that the performance is far worse when the model translates rare words.

Since the pre-trained monolingual word embeddings provide the cornerstone for learning unsupervised word mapping, the quality of monolingual embeddings directly determines the quality of word mapping. Due to the low frequency of rare words, the quality of their embeddings is lower than that of common words. This makes the isometric assumption [6] more difficult to satisfy on rare words, leading to poor performance of all methods on rare word pairs. Improving the quality of cross-lingual embeddings of rare words is expected to be explored in future work.

6 Related Work

This paper is mainly related to the following two lines of work.

Supervised cross-lingual embedding. Inspired by the isometric observation between monolingual word embeddings of two different languages, [19] proposes to learn the desired word mapping by minimizing mean squared error. At the inference stage, they

adopt cosine similarity as the distance metric to fetch the translation of a word. Furthermore, [9] investigates the hubness problem and [10] incorporates the semantics of a word in multiple languages into its embedding. [23] argues that the results can be improved by imposing the orthogonal constraint to the linear mapping. There also exist some other representative researches. For instance, [22] presents inverse-softmax which normalizes the softmax probability over source words rather than target words and [4] presents a self-learning framework to perform iterative refinement.

Unsupervised cross-lingual embedding. The endeavors to explore unsupervised cross-lingual embedding are mainly divided into two categories. One line of research focuses on designing heuristics or utilizing the structural similarity of monolingual embeddings. For instance, [14] presents a non-adversarial method based on the principal component analysis. Both [1] and [6] take advantage of geometric properties across languages to perform word retrieval to learn the initial word mapping. However, these methods usually require plenty of random restarts or additional skills to achieve satisfactory performance. Another line strives to learn unsupervised word mapping by directly perform distribution-matching. For example, [17] and [25] completely eliminate the need for any supervision signal by aligning the distribution of transferred embedding and target embedding with GAN. [26] and [24] adopt the Earth Mover's distance and Sinkhorn distance as the optimized distance metrics respectively, which requires intermediate *density estimation*. Although this line achieves relatively excellent performance, they suffer from a sophisticated alternate optimization, which tends to cause a hard and unstable training process. There are also some attempts to improve distant language pairs. For instance, [15] generalizes Procrustes analysis by projecting the two languages into a latent space and [20] proposed to learn neighborhood sensitive mapping by training non-linear functions.

7 Conclusion

In this paper, we propose to learn unsupervised word mapping between different languages by directly minimize the maximum mean discrepancy between the distribution of transferred embedding and target embedding. The proposed model adopts non-parametric metric that does not require any intermediate *density estimation* or adversarial training. This avoids a relatively sophisticated and unstable alternate optimization process. Experimental results show that the proposed method can achieve better performance than several state-of-the-art systems. Further analysis demonstrates the effectiveness of our approach in improving stability.

References

1. Aldarmaki, H., Mohan, M., Diab, M.T.: Unsupervised word mapping using structural similarities in monolingual embeddings. TACL **6**, 185–196 (2018)
2. Alvarez-Melis, D., Jaakkola, T.S.: Gromov-wasserstein alignment of word embedding spaces. In: EMNLP. pp. 1881–1890 (2018)

3. Artetxe, M., Labaka, G., Agirre, E.: Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In: EMNLP. pp. 2289–2294 (2016)
4. Artetxe, M., Labaka, G., Agirre, E.: Learning bilingual word embeddings with (almost) no bilingual data. In: ACL. pp. 451–462 (2017)
5. Artetxe, M., Labaka, G., Agirre, E.: Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations. In: AAAI. pp. 5012–5019 (2018)
6. Artetxe, M., Labaka, G., Agirre, E.: A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In: ACL. pp. 789–798 (2018)
7. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. TACL **5**, 135–146 (2017)
8. Chen, X., Cardie, C.: Unsupervised multilingual word embeddings. In: EMNLP. pp. 261–270 (2018)
9. Dinu, G., Baroni, M.: Improving zero-shot learning by mitigating the hubness problem. In: ICLR (2015)
10. Faruqui, M., Dyer, C.: Improving vector space word representations using multilingual correlation. In: EACL. pp. 462–471 (2014)
11. Grave, E., Joulin, A., Berthet, Q.: Unsupervised alignment of embeddings with wasserstein procrustes. arXiv:1805.11222 (2018)
12. Gretton, A., Borgwardt, K., Rasch, M., Schölkopf, B.: A kernel method for the two-sample-problem. In: NIPS. pp. 513–520 (2006)
13. Gretton, A., Borgwardt, K.M., Rasch, M.J., Schölkopf, B., Smola, A.: A kernel two-sample test. Journal of Machine Learning Research **13**, 723–773 (2012)
14. Hoshen, Y., Wolf, L.: An iterative closest point method for unsupervised word translation. arXiv:1801.06126 (2018)
15. Kementchedjhieva, Y., Ruder, S., Cotterell, R., Søgaard, A.: Generalizing procrustes analysis for better bilingual dictionary induction. In: CoNLL. pp. 211–220 (2018)
16. Kondrak, G., Hauer, B., Nicolai, G.: Bootstrapping unsupervised bilingual lexicon induction. In: EACL. pp. 619–624 (2017)
17. Lample, G., Conneau, A., Ranzato, M., Denoyer, L., Jégou, H.: Word translation without parallel data. In: ICLR (2018)
18. Li, Y., Swersky, K., Zemel, R.S.: Generative moment matching networks. In: ICML. pp. 1718–1727 (2015)
19. Mikolov, T., Le, Q.V., Sutskever, I.: Exploiting similarities among languages for machine translation. arXiv:1309.4168 (2013)
20. Nakashole, N.: NORMA: neighborhood sensitive maps for multilingual word embeddings. In: EMNLP. pp. 512–522 (2018)
21. Shigeto, Y., Suzuki, I., Hara, K., Shimbo, M., Matsumoto, Y.: Ridge regression, hubness, and zero-shot learning. In: PKDD. pp. 135–151 (2015)
22. Smith, S.L., Turban, D.H., Hamblin, S., Hammerla, N.Y.: Offline bilingual word vectors, orthogonal transformations and the inverted softmax. arXiv:1702.03859 (2017)
23. Xing, C., Wang, D., Liu, C., Lin, Y.: Normalized word embedding and orthogonal transform for bilingual word translation. In: NAACL. pp. 1006–1011 (2015)
24. Xu, R., Yang, Y., Otani, N., Wu, Y.: Unsupervised cross-lingual transfer of word embedding spaces. arXiv:1809.03633 (2018)
25. Zhang, M., Liu, Y., Luan, H., Sun, M.: Adversarial training for unsupervised bilingual lexicon induction. In: ACL. pp. 1959–1970 (2017)
26. Zhang, M., Liu, Y., Luan, H., Sun, M.: Earth mover's distance minimization for unsupervised bilingual lexicon induction. In: EMNLP. pp. 1934–1945 (2017)
27. Zhang, Y., Gaddy, D., Barzilay, R., Jaakkola, T.S.: Ten pairs to tag - multilingual POS tagging via coarse mapping between embeddings. In: NAACL. pp. 1307–1317 (2016)